NLP ANALYSIS ON REDDIT

by. Suchanya (Mild) Trakarnsakdikul

How I Met Your Mother VS The Big Bang Theory

Have you heard about...

OR

What is it about?

The Problem

Identify what keywords are the best to used for searching about the selected shows.

The Data

How I Met Your Mother

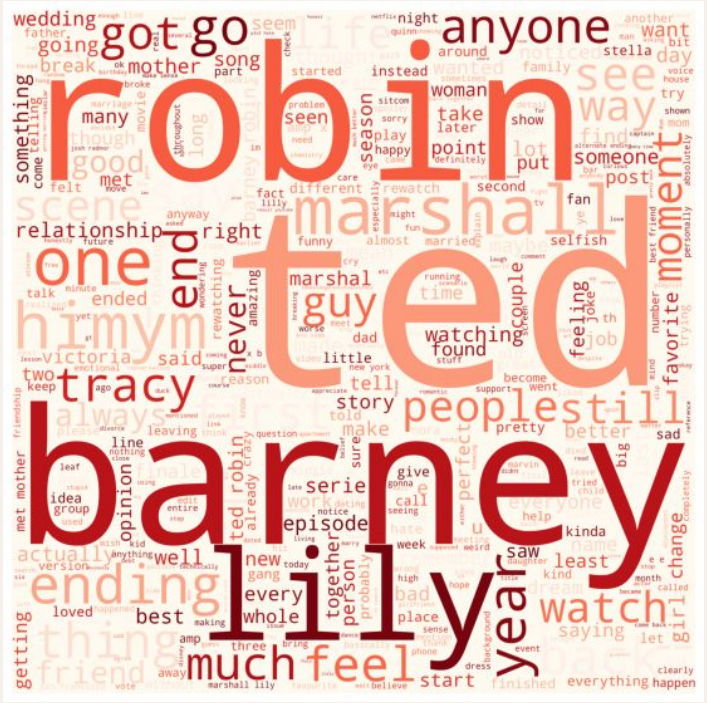
- 869 reddit posts
- Consisted of 114 features
- Grouped into subreddit, text, media
- Show set in New Your City
- Ran from 2005 to 2014

The Big Bang Theory

- 1012 reddit posts
- Consisted of 114 features
- Grouped into subreddit, text, media
- Show set in California
- Ran from 2007 to 2019

A Quick Glance

How I Met Your Mother



VS

The Big Bang Theory

Modeling and Prediction

Logistic, KNN and Naïve Bayes Estimators

- Count Vectorized to determined the value of each words
- Fitted the best parameters for the estimators and vectorizers
- From the quality parameters an optimized model is created with high accuracy

Best Model Evaluation

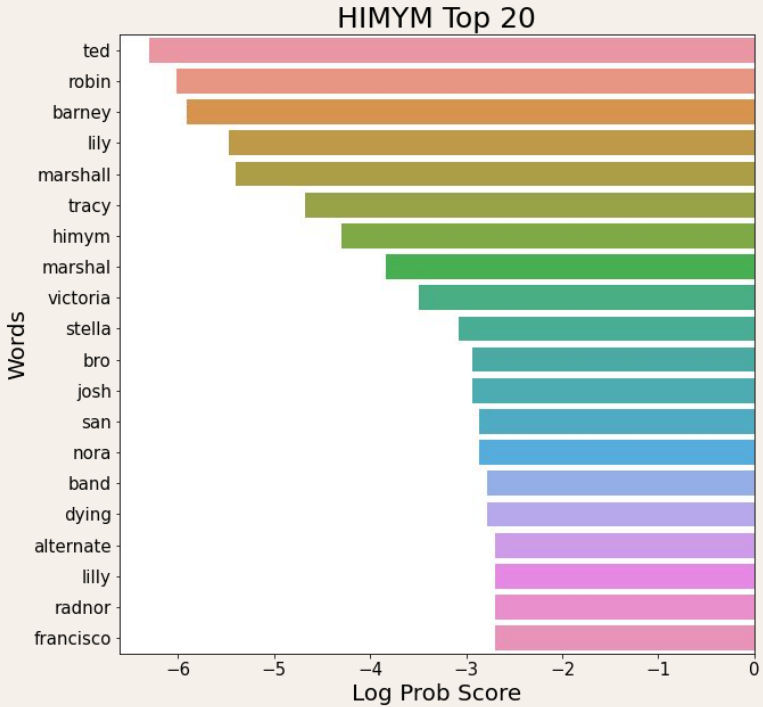
The Naïve Bayes estimators gave 86% accuracy rate

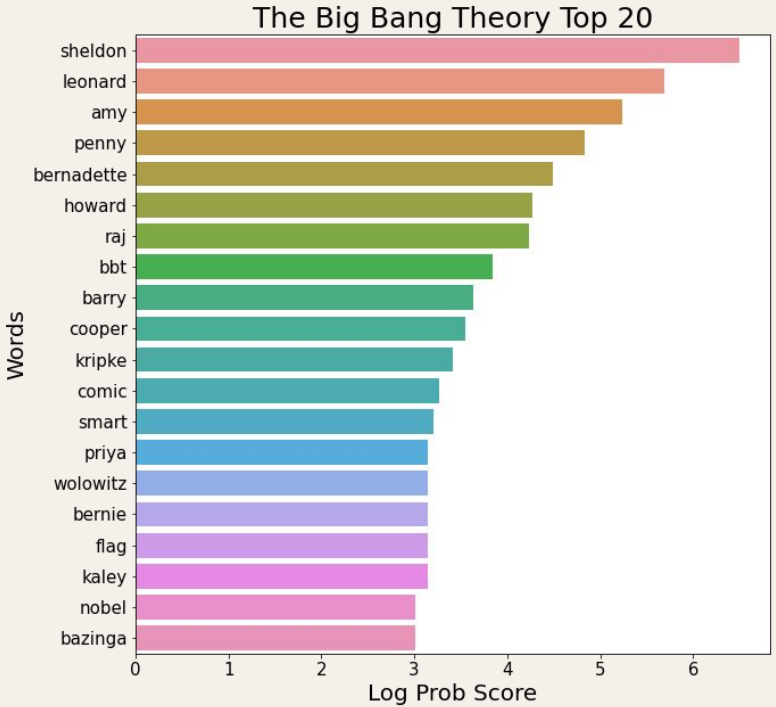
- Used all text data from titles and post contents
- 84% of the posts were correctly classified as HIMYM posts
- 87% of HIMYM posts are classified correctly
- 85% of The Big Bang Theory posts are classified correctly

	predict Big Bang Theory	predict HIMYM
actual Big Bang Theory	168	29
actual HIMYM	22	158

```
Accuracy: 0.8647
Misclassification rate: 0.1353
Precision: 0.8449
Recall: 0.8778
Specificity: 0.8528
```

Keyword Recommendation





Additional Data

From log probability evaluation

- Positive EvalScore suggest post leans towards HIMYM
- Negative EvalScore suggest post leans towards Big Bang Theory
- Post closer to 0 should be remove as it can influence the prediction
- Gain more data than the 1888 that were used to model

	Word	EvalScore
3178	non	0.040079
909	compared	0.040079
155	amanda	0.040079
96	age	0.040079
4037	rude	0.040079
4718	teen	0.040079
3496	playing	0.040079
2790	long	0.040079
1594	every	0.040079
1624	excuse	0.040079
257	arguably	0.040079
4847	totally	0.040079
3632	probably	0.040079
436	become	0.040079
1692	famous	0.040079
4741	terrible	0.040079
2930	mean	0.040079
4779	thought	0.061725
1860	found	0.103191
1028	could	0.107978

	Word	EvalScore
541	bob	-2.218704
4515	step	-2.218704
5101	vote	-2.046689
5202	wide	-1.624929
2810	loud	-1.624929
4118	scream	-1.624929
1665	eye	-1.336913
1224	depth	-1.309848
1783	fish	-1.309848
648	butter	-1.309848
3891	release	-1.309848
2781	lockdown	-1.309848
3556	post	-1.240133
838	clip	-1.169759
1266	die	-1.169759
996	conversation	-1.129992
2765	literally	-1.100206
584	break	-1.093020
161	amazing	-1.029755
5306	youtube	-1.021832

Thank you for listening!

Appreciate your time